Resumen

La detección del lenguaje abusivo es una tarea que se ha vuelto cada vez más importante en la era digital moderna, donde la comunicación se produce a través de diversas plataformas en línea. El aumento de las interacciones en estas plataformas ha provocado un aumento de la aparición del lenguaje abusivo. Abordar dicho contenido es crucial para mantener un entorno en línea seguro e inclusivo. Sin embargo, esta tarea enfrenta varios desafíos que la convierten en un área compleja y que demanda de continua investigación y desarrollo. En particular, detectar lenguaje abusivo en entornos con escasez de datos presenta desafíos adicionales debido a que el desarrollo de sistemas automáticos precisos a menudo requiere de grandes conjuntos de datos anotados.

En esta tesis investigamos diferentes aspectos de la detección del lenguaje abusivo, prestando especial atención a entornos con datos limitados. Primero, estudiamos el sesgo hacia palabras clave abusivas en modelos entrenados para la detección del lenguaje abusivo. Con este propósito, proponemos dos métodos para extraer palabras clave potencialmente abusivas de colecciones de textos. Luego evaluamos el sesgo hacia las palabras clave extraídas y cómo se puede modificar este sesgo para influir en el rendimiento de la detección del lenguaje abusivo. El análisis y las conclusiones de este trabajo revelan evidencia de que es posible mitigar el sesgo y que dicha reducción puede afectar positivamente el desempeño de los modelos. Sin embargo, notamos que no es posible establecer una correspondencia similar entre la variación del sesgo y el desempeño de los modelos cuando hay escasez datos con las técnicas de reducción del sesgo estudiadas.

En segundo lugar, investigamos el uso de redes neuronales basadas en grafos para detectar lenguaje abusivo. Por un lado, proponemos una estrategia de representación de textos diseñada con el objetivo de obtener un espacio de representación en el que los textos abusivos puedan distinguirse fácilmente de otros textos. Por otro lado, evaluamos la capacidad de redes neuronales convolucionales basadas en grafos para clasificar textos abusivos. La siguiente parte de nuestra investigación se centra en analizar cómo el aumento de datos puede influir en el rendimiento de la detección del lenguaje abusivo. Para ello, investigamos dos técnicas bien conocidas basadas en el principio de minimización del riesgo en la vecindad de instancias originales y proponemos una variante para una de ellas. Además, evaluamos técnicas simples basadas en el reemplazo de sinónimos, inserción aleatoria, intercambio aleatorio y eliminación aleatoria de palabras. Las contribuciones de esta tesis ponen de manifiesto el potencial de las redes neuronales basadas en grafos y de las técnicas de aumento de datos para mejorar la detección del lenguaje abusivo, especialmente cuando hay limitación de datos. Estas contribuciones han sido publicadas en conferencias y revistas internacionales.